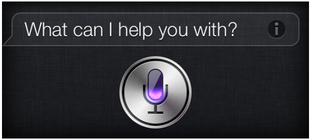
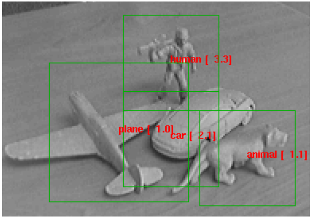
Multilayer Networks

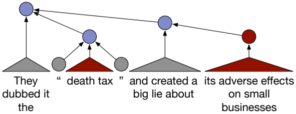
Machine Learning: Jordan Boyd-Graber
University of Maryland
INTRODUCTION

Deep Learning was once known as “Neural Networks”

But it came back . . .







- More data
- Better tricks (regularization)
- Faster computers

And companies are investing . . .

Google Hires Brains that Helped Supercharge Machine Learning

BY ROBERT MCMILLAN 03.13.13 | 6:30 AM | PERMALINK

Share 0 Tweet 1 +1 145 Share Pin it

And companies are investing . . .

‘Chinese Google’ Opens Artificial-Intelligence Lab in Silicon Valley

BY DANIELA HERNANDEZ 04.12.13 | 6:30 AM | PERMALINK

Share 0 Tweet 0 +1 228 Share Pin it

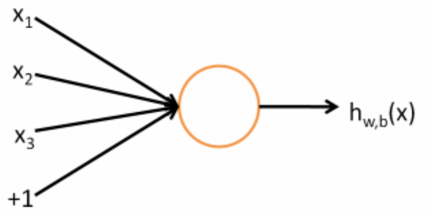
And companies are investing . . .

Facebook's 'Deep Learning' Guru Reveals the Future of AI

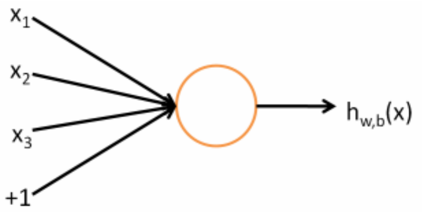
BY CADE METZ 12.12.13 | 6:30 AM | PERMALINK

Share 2 Tweet 1 +1 143 Share Pin it

Map inputs to output



Map inputs to output

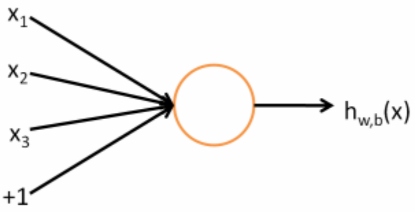


Input

Vector $x_1 \dots x_d$

inputs encoded as real numbers

Map inputs to output



Input

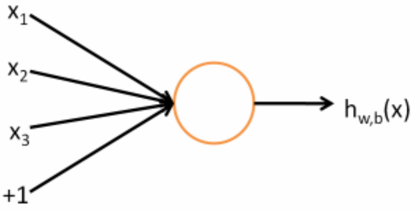
Vector $x_1 \dots x_d$

Output

$$f\left(\sum_i W_i x_i + b\right)$$

multiply inputs by

Map inputs to output



Input

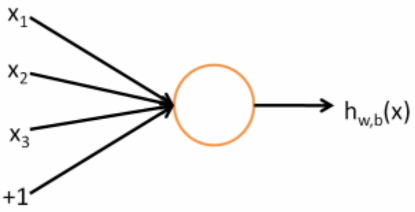
Vector $x_1 \dots x_d$

Output

$$f\left(\sum_i W_i x_i + b\right)$$

add bias

Map inputs to output



Input

Vector $x_1 \dots x_d$

Output

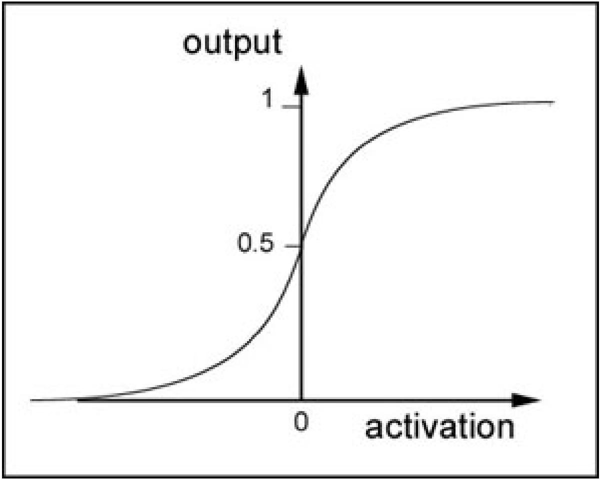
$$f\left(\sum_i W_i x_i + b\right)$$

Activation

$$f(z) \equiv \frac{1}{1 + \exp(-z)}$$

pass through nonlinear sigmoid

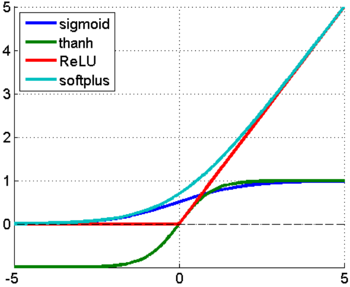
Why is it called activation?



In the shallow end

- This is still logistic regression
- Engineering features x is difficult (and requires expertise)
- Can we learn how to represent inputs into final decision?

Better name: non-linearity



- Logistic / Sigmoid

$$f(x) = \frac{1}{1+e^{-x}} \tag{1}$$

- tanh

$$f(x) = \tanh(x) = \frac{2}{1+e^{-2x}} - 1 \tag{2}$$

- ReLU

$$f(x) = \begin{cases} 0 & \text{for} \quad x < 0 \\ x & \text{for} \quad x \geq 0 \end{cases} \tag{3}$$

- SoftPlus: $f(x) = \ln(1+e^{x})$

But it is not perfect

- Compare against baselines: randomized features, nearest-neighbors, linear models
- Optimization is hard (alchemy)
- Models are often not interpretable
- Requires specialized hardware and tons of data to scale